DOCUMENT RESUME

ED 204 401 TM 810 415

AUTHOR Qualls, Audrey L.
TITLE Black and White Teacher Ratings of Elementary Achievement Test Items for Potential Race Favoritism.
PUB DATE Apr 81
NOTE 21p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests: Blacks: Elementary Education: *Standardized Tests: *Teacher Attitudes: *Test Bias: Whites
IDENTIFIERS Chicago Public Schools IL: *Iowa Tests of Basic Skills

ABSTRACT

Teacher perceptions of standardized achievement test items, drawn mainly from the Iowa Tests of Basic Skills (ITBS), as possibly favoring black or white pupils was investigated. Teacher' ratings were compared across race of raters, black and white: item types: favoring blacks and favoring whites: and six subtests. Ratings were also compared with two statistical indices to ascertain their relationship to common item bias identification procedures. Thiry items with the most extreme black-white differences in transformed item difficulty from the vocabulary, spelling, capitalization, punctuation, and language usage subtests (a total pool of 453 items) were selected for teacher evaluation. Eight additional items labeled Intentional Favoritism (IF) appeared to favor either black or white pupils. The sample consisted of 41 white and 51 black teachers selected from one school district in Chicago and three additional public elementary Chicago schools. Results implied that teacher reviewers as a group can reliably and validly rate test items for favoritism in either the direction of blacks and whites as indicated by their ratings on the IF. In particular, the limited effects for item type differences on the 30 ITBS items can be attributed to editorial scrutiny for both ambiguities and bias. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED204401

BLACK AND WHITE TEACHER RATINGS OF ELEMENTARY ACHIEVEMENT TEST ITEMS FOR POTENTIAL RACE FAVORITISM

Audrey L. Qualls

Dallas Independent School District

H. D. Hoover

University of Iowa

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. Qualls

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 415

Paper presented at the Annual Meeting of the American Educational Research Association

Los Angeles, April 1981

Introduction

Bias in an achievement test is usually assessed through the statistical inspection of relationships within the test at the individual item level. This investigation is usually of interest during the test construction phase, so that identified items may be edited before use or deleted from the test. The investigation of bias in achievement testing is not limited to statistical methods. Test developers rely heavily upon the opinions of "experts". Trained professionals from minority groups review the test material with an eye for bias, content specialists determine what will be tested, and editors review the items to eliminate possible ambiguities (Merz, 1978; Green, 1978). Even with these efforts, systematic differences on test performance, which often appear to be unrelated to examinee ability, do occur. At the moment, there are no procedures available that can guarantee elimination of biased items in achievement tests. Researchers must continue to explore different avenues and to refine present procedures.

A professional group of individuals frequently ignored in the search for different item bias detection methods is teachers. Teachers interact daily with their students, thus allowing them to form expectations of student abilities based on many observations and evaluations. The knowledge regarding student backgrounds and abilities that teachers accumulate on a daily basis may be useful in identifying biased items. Teacher contributions to resolving the problem of test bias in standardized tests should be explored.

Purpose of the Study

The purpose of this study was to investigate teacher perceptions of standardized achievement test items as possibly favoring black or white pupils. The teachers' ratings were compared across race of raters, black and white; item types, favoring blacks and favoring whites; and six subtests. The ratings were also compared with two statistical indices to ascertain their relationship to common item bias identification procedures. The major intent of such comparisons is to determine whether teachers' ratings vary and interact with teacher race, item types, and subtests.

Procedure

Description of the Instrument

Items for this study were drawn mainly from the Iowa Tests of Basic Skills (ITBS). In the most recent national standardization of the ITBS, data were obtained from representative subsamples of 400 black and 400 white pupils. A transformed item difficulty procedure was conducted to identify items that tended to be relatively easier for white pupils as a group or black pupils as a group. The 30 items with the most extreme black-white differences in transformed item difficulty from the Vocabulary, Spelling, Capitalization, Punctuation, and Language Usage subtests (a total pool of 453 items) were selected for teacher evaluation in this study. Items extremely difficult for both groups were excluded from consideration. Eight additional items labeled Intentional Favoritism were written by the first author. The intent was to write appropriate Vocabulary and Language Skills items that obviously appeared to favor either black or white pupils. It should be noted that the teachers were unaware of which items were specifically written for this

study. A copy of the evaluation form containing both the directions and items rated appears in the Appendix. The (B) and (W) entered by hand beneath the item numbers and intended age levels for the ITBS items indicate the group relatively favored on the basis of the obtained differences in transformed item difficulty. The items intentionally written to favor either blacks or whites are boxed and the group favored is indicated. This information was not available to the teacher raters.

Sample

The sample consisted of 41 white and 51 black teachers selected from one school district in Chicago and three additional public elementary Chicago schools. All were teachers in an inner city area with student populations greater than 90 per cent black.

Data Analysis

The analysis procedures used in this study were designed to provide answers to the following main questions:

1. Can black and white teachers identify items that statistically favor white pupils as a group and/or black pupils as a group across subtests?

2 To what extent, if any, does teacher race interact with ratings of items favoring blacks and items favoring whites?

3. To what extent do teacher ratings correlate with statistical indices of bias: differences in item difficulties and differences in item discrimination (point biserials).

The first step in data analysis involved an analysis of variance (ANOVA) procedure. A three-factor ANOVA design with repeated measures on two factors was employed for the purpose of investigating the effect of item type upon subtests ratings across teacher race. The three factors

were: (a) teacher race, which had two levels, blacks and whites; (b) item type (direction of favoritism), which had two levels, favoring whites and favoring blacks; (c) subtest, which had six levels, Vocabulary (V), Spelling (L-1), Capitalization (L-2), Punctuation (L-3), Usage (L-4), and Intentional Favoritism (IF).

The cell scores were an average of all ratings within a subtest for a particular item type and subject. The averages for each individual within a cell for all subtests, excluding IF, were a result of three ratings. Averages for the IF subtest were based on four ratings. The scale on which the teacher ratings were obtained ranged from 1 to 5 with a midpoint of 3. For the purposes of scoring the scale was transformed to a scale ranging from 1 to 9 with a midpoint of 5. One (1) indicated definite favoritism for whites, 5 denoted neutrality, and 9 indicated definite favoritism for blacks.

Results

General

Prior to the consideration of the results from the ANOVA, intraclass reliability analyses were conducted. The results indicated a reliability coefficient for the average ratings, on the 38 items of .27. Table 1 presents the intraclass reliabilities for each subtest and teacher group separately and together. It is of interest to note that the only "sizeable" reliability coefficient for teacher ratings on an ITBS subtest was found on Black teachers' ratings for capitalization items. Closer inspection of the data indicated that the .33 coefficient was primarily

TABLE 1

Intraclass Reliabilities of Ratings

SUBTEST	White Teachers N=41	Black Teachers N=51	Combined N=92
Vocabulary	.03	.16	.09
Spelling	.05	.07	.05
Capitalization	.03	.33	.18
Punctuation	.07	.04	.04
Usage	.01	.00	.00
Intentional Favoritism	.39	.60	.51
30 ITBS items	.03	.11	.18
38 items	.20	.34	.27

the result of one item, number 17. Black teachers as a group, more so than whites, felt that this item favored whites; whereas pupil data indicated that item number 17 relatively favored blacks. It should be noted that the reliabilities for the IF subtest were by far the largest. The overall results from the intraclass reliability analyses lead to the conclusion that individual teacher's ratings, like a single test item, are quite unreliable. For the 38 items, the reliability of the mean item rating for all 92 teachers was .97.

Effects for ANOVA

The expected double interaction of item type by subtest was obtained. This interaction is shown plotted in Figure 1. The results indicate that the mean differences in teacher ratings on items favoring whites and items favoring blacks are different for the various subtests.

There was no significant interaction for teacher race by item type or teacher race by subtests. A perusal of Table 2 indicates that black and white teachers' marginal item type ratings and subtest ratings are similar.

A significant triple interaction for teacher race by item type by subtest was also identified. Follow-up analyses were concerned with examining the overall differences in teacher ratings for blacks and whites on two item types for each individual subtest. Results from the six subtest ANOVA's identified three significant race by item type interactions: Spelling, Capitalization, and Intentional Favoritism. The pattern of two-way interactions were not always the same. On the Spelling test, white teachers rated items favoring whites "higher" than items favoring blacks, while black teachers rated items favoring whites "lower" than items favoring blacks. "Higher" ratings indicate that the

Figure 1

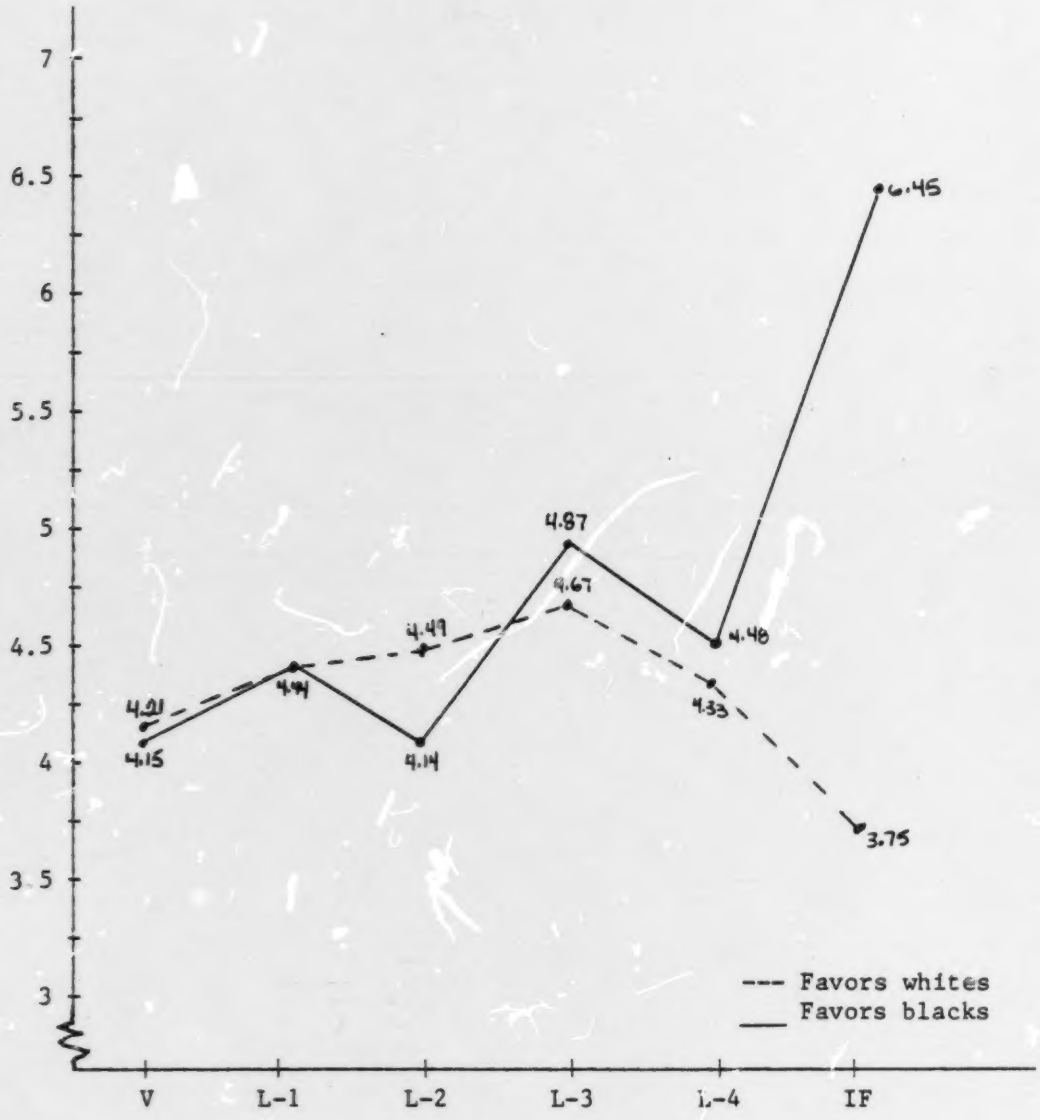


Figure . Double interaction of item types across race by subtest.

Table 2

Cell Mean Ratings, Standard Deviations and Marginal Means

TEACHER RACE / SUBTEST	ITEM TYPES: Favoring Whites — White	Favoring Whites — Blacks	Favoring Blacks — White	Favoring Blacks — Blacks	ACROSS ITEM TYPES — White Teachers	Black Teachers
V						
$\bar{X}$	4.32	4.12	4.12	4.17	4.22	4.14
S.D.	.66	.74	.71	.87		
$\bar{X}$ Both races	4.21		4.15		4.18	
L-1						
$\bar{X}$	4.56	4.35	4.32	4.53	4.44	4.44
S.D.	.75	.78	.71	1.04		
$\bar{X}$ Both races	4.44		4.44		4.44	
L-2						
$\bar{X}$	4.50	4.48	4.37	3.95	4.43	4.21
S.D.	.74	.58	.63	.74		
$\bar{X}$ Both races	4.49		4.14		4.31	
L-3						
$\bar{X}$	4.72	4.66	4.80	4.93	4.76	4.79
S.D.	.63	.62	.63	.49		
$\bar{X}$ Both races	4.67		4.87		4.78	
L-4						
$\bar{X}$	4.38	4.29	4.71	4.29	4.54	4.29
S.D.	1.70	1.40	1.46	1.30		
$\bar{X}$ Both races	4.33		4.48		4.40	
IF						
$\bar{X}$	3.91	3.55	6.16	6.64	5.03	5.09
S.D.	.93	.95	.93	.83		
$\bar{X}$ Both races	3.75		6.45		5.06	
30 ITBS $\bar{X}$	4.50	4.38	4.46	4.37	4.48	4.37
	4.43		4.41		4.42	
Overall $\bar{X}$	4.37	4.20	4.82	4.85	4.59	4.52
Item Type $\bar{X}$	4.28		4.84		4.55	

item favors blacks and "lower" ratings indicate relative favoritism toward whites. For Capitalization, both teacher groups rated white items virtually the same, but items favoring blacks were rated relatively higher by white teachers. The Intentional Favoritism item ratings concurred with a priori beliefs; they were rated in te expected directions and markedly different. Black teachers' mean ratings on both item types indicated "larger" degrees of favoritism, which could suggest a greater sensitivity toward cultural differences.

A significant racial main effect was found on the Capitalization test, which lead to the conclusion that item type ratings do differ on this test depending on the race of the rater. Whites tended to rate both item types slightly higher.

As expected, within subtests there were main effects for item types but only on the Capitalization, Punctuation, and Intentional Favoritism subtests. However, the item ratings do differ by subtest. Item type ratings on the Punctuation and Intentional Favoritism subtests were lower on items favoring whites and higher on items favoring blacks. These results were in agreement with a priori beliefs and the initial arcsin item difficulty differences.

Correlations between Teacher Ratings and Statistical Indices

The mean teacher ratings for each of 30 ITBS items per racial group were compared to two statistical item bias detection procedures; item difficulty and discrimination differences. The correlations among these procedures are presented in Table 3. The results were a bit surprising, but conclusions were inappropriate prior to testing the significance of these correlations. Two-tailed t-tests were performed to determine the significance of each of these correlations. These tests indicated

Table 3

Correlational Matrix for
Item Difficulties, Discriminations, and Teacher Ratings

	Difficulty Differences	Discrimination Differences	White Teachers	Black Teachers
Difficulty Differences				
Discrimination Differences	-.15			
White teacher's Ratings	-.10	-.14		
Black teacher's Ratings	-.11	-.21	.63	

nonsignificant results for all, suggesting a lack of relationship between teacher ratings and these two item bias detection procedures.

Implications of the Findings

The results of this study imply that teacher reviewers as a group can reliably and validly rate test items for favoritism in either the direction of blacks and whites as indicated by their ratings on the Intentional Favoritism test.

In particular, the limited effects for item type differences on the 30 ITBS items can be attributed to editorial scrutiny for both ambiguities and bias. Editorial reviews seem essential for reducing test performance differences arising from pupil background variables rather than ability. Items on the IF test, which were written to favor either black or white pupils, were subject to editorial review for grammatical errors only. The marked differences in ratings for this test, in comparison to the ITBS subtests, is probably the result of editorial work on ITBS items.

It is recognized that this study does not provide sufficient evidence to prove, disprove or evaluate the methods used for item bias identification, nor to develop a more precise approach to the study of bias. It does, however, present an additional alternative, teacher reviewers which warrants further investigation.

APPENDIX

A Fairness Evaluation of Selected Vocabulary
ard Language Skill Items

The purpose of this research project is to determine whether it is possible to use teacher ratings to identify test items that function differently for black children and for white children. In the most recent national standardization of the Iowa Tests of Basic Skills (ITBS), data were obtained from representative subsamples of black and white pupils. On the basis of these data, items were selected through statistical means that tended to be relatively easier for white pupils as a group or for black pupils as a group, perhaps because of the children's ethnic background and/or other characteristics.

The purpose of this survey is to obtain from experienced teachers ratings of the apparent fairness of selected items. The teachers' ratings will be compared with the statistical indices to determine to what extent teacher ratings and statistical procedures identify the same items as possibly biased in favor of white children or black children.

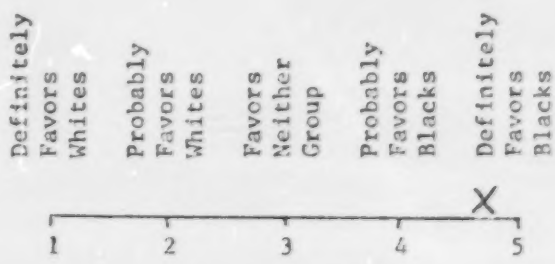
DIRECTIONS
You are asked to rate 38 items on a five-point scale, based on your perception of their fairness to black and white pupils of comparable ability. The 38 items are a combination of actual ITBS items and items written especially for this study.

The vocabulary items require pupils to identify synonyms. The language skill items require pupils to identify the line (if any) containing mistakes in spelling, capitalization, punctuation, and "standard" English usage.

The following two examples may aid you in this task:

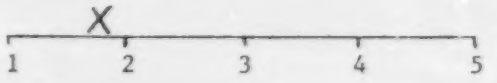
Capitalization

1. (11-12)
1) Audrey's mother decided that the
2) best birthday gift for her daughter
(3) would be a subscription to ebony jr.
4) (no mistakes)



Usage

2. (9-10)
1) Jesse is stubborn. Linda told him
(2) that Michael done gone to the store,
3) but he refused to believe her.
4) (no mistakes)



The correct answers are circled and the intended age level of each item is presented in parentheses below the item number.

In addition to rating each item, please check the box to the right of each rating scale if, in your opinion, the item should not be included on an achievement test at this age level. The fact that an item favors one group does not always mean it should be deleted from a test. Some items that seem to favor one group may be important measures of a pupil's educational development.

Please supply the following biographical information:

Age:_____ Sex:_____ Race:_____

Years teaching experience:_____ Age group of pupils:_____

VOCABULARY

	Definitely Favors Whites	Probably Favors Whites	Favors Neither Group	Probably Favors Blacks	Definitely Favors Blacks	Item Should Not Be Included
1. Underneath the chair (9) ① below W 2) around 3) near 4) throne	1	2	3	4	5	☐
2. A brown couch (9) ① sofa B 2) chair 3) table 4) throne	1	2	3	4	5	☐
3. Invent an excuse (9-10) 1) believe B 2) explain 3) accept ④ make up	1	2	3	4	5	☐
4. To buy in quantity (9-10) 1) installments B ② large amounts 3) season 4) pieces	1	2	3	4	5	☐
5. The object of the game (11-13) 1) rules W ② purpose 3) story 4) ending	1	2	3	4	5	☐
6. A victorious team (12-14) 1) promising W 2) fortunate 3) powerful ④ triumphant	1	2	3	4	5	☐
7. The immense forest (13-14) 1) dark B ② huge 3) dense 4) ancient	1	2	3	4	5	☐
8. A gospel choir (13-14) 1) true B ② religious 3) social 4) accapella	1	2	3	4	5	☐

LANGUAGE SKILLS L-1: SPELLING

			Definitely Favors Whites	Probably Favors Whites	Favors Neither Group	Probably Favors Blacks	Definitely Favors Blacks	Item Should Not Be Included
9. (9-10) W	(1)	clath						
	2)	note						
	3)	depart						
	4)	windy						
	5)	(No mistakes)						
10. (10-11) W	1)	tale						
	2)	oars						
	3)	rate						
	4)	germ						
	(5)	(No mistakes)						
11. (10-12) B	1)	cannot						
	2)	bedroom						
	(3)	adition						
	4)	hollow						
	5)	(No mistakes)						
12. (11-12) B	(1)	retern						
	2)	permit						
	3)	melted						
	4)	subject						
	5)	(No mistakes)						
13. (11-12) W	1)	comma						
	2)	savage						
	3)	coward						
	(4)	iland						
	5)	(No mistakes)						
14. (13-14) B	1)	rotation						
	(2)	elaborete						
	3)	completion						
	4)	patronize						
	5)	(No mistakes)						

LANGUAGE SKILLS L-2: CAPITALIZATION

			Definitely Favors Whites	Probably Favors Whites	Favors Neither Group	Probably Favors Blacks	Definitely Favors Blacks	Item Should Not Be Included
15. (9-10) W	1) 2) 3) (4)	My mother and I walked to church along the path that runs by the creek. (No mistakes)	1	2	3	4	5	☐
16. (9-10) B	1) (2) 3) 4)	The girls were playing Double Dutch with the rope Karen's mother gave her. (No mistakes)	1	2	3	4	5	☐
17. (9-11) B	(1) 2) 3) 4)	The part of peter in the play *Peter Pan* is often played by a woman. (No mistakes)	1	2	3	4	5	☐
18. (9-11) B	1) (2) 3) 4)	Beth knew that the first Monday in September was labor day, but she forgot it was a holiday. (No mistakes)	1	2	3	4	5	☐
19. (9-11) W	1) 2) (3) 4)	6 Oak Avenue Twin Falls, ID 83301 june 20, 1977 (No mistakes)	1	2	3	4	5	☐
20. (10-11) W	1) 2) (3) 4)	Maurice slid into second base and tore his brand-new pair of Dungaree pants. (No mistakes)	1	2	3	4	5	☐
21. (11-13) W	1) 2) (3) 4)	When the economy is down, the sales of duncan yo-yos go up. (No mistakes)	1	2	3	4	5	☐
22. (11-13) B	1) 2) (3) 4)	The steeple of Old North Church in Boston was rebuilt through children's donations after a 1954 Hurricane. (No mistakes)	1	2	3	4	5	☐
23. (14) W	1) 2) (3) 4)	Mendelssohn was the German composer who wrote the overture to A Midsummer night's dream. (No mistakes)	1	2	3	4	5	☐

LANGUAGE SKILLS L-3: PUNCTUATION

			Definitely Favors Whites	Probably Favors Whites	Favors Neither Group	Probably Favors Blacks	Definitely Favors Blacks	Item Should Not Be Included
24. (9-10) W	1) 2) 3) (4)	It took Karen two hours to teach Barkus to jump through a hoop. (No mistakes)	1	2	3	4	5	☐
25. (9-10) B	1) (2) 3) 4)	The discovery of gold on January 24 1848 started the California Gold Rush. (No mistakes)	1	2	3	4	5	☐
26. (10-11) B	1) 2) (3) 4)	Carolyn's vegetable garden was a mess. The weeds had outgrown the collards mustards and cabbage. (no mistakes)	1	2	3	4	5	☐
27. (10-12) B	1) (2) 3) 4)	Three coats of paint didn t cover all the scratches on the table. (No mistakes)	1	2	3	4	5	☐
28. (10-12) B	1) (2) 3) 4)	The first U.S. coins were minted in Philadelphia Pennsylvania, then the nation's capital. (No mistakes)	1	2	3	4	5	☐
29. (13-14) W	(1) 2) 3) 4)	Eyeglasses counter tops and toys are now safer and stronger, because they are made from plastic. (No mistakes)	1	2	3	4	5	☐
30. (14) W	(1) 2) 3) 4)	The fish except for two were very small. The two big ones were the ones Dad caught. (No mistakes)	1	2	3	4	5	☐

LANGUAGE SKILLS 1-4: USAGE

			Definitely Favors Whites	Probably Favors Whites	Favors Neither Group	Probably Favors Blacks	Definitely Favors Blacks	Item Should Not Be Included
31. (9-10) B	1) 2) ③ 4)	Ben has fat fingers. Once he got one stuck in a phone. The dial had to be tooken off. (No mistakes)	1	2	3	4	5	☐
32. (9-10) W	1) 2) ③ 4)	Birds are smart. They do things together. They all goes south in winter. (No mistakes)	1	2	3	4	5	☐
33. (10-11) W	1) 2) ③ 4)	Cheryl was supposed to meet Sam in the schoolyard. By the time she got there he been gone. (No mistakes)	1	2	3	4	5	☐
34. (10-12) W	① 2) 3) 4)	A frog likes food that move. He won't even touch a dead fly. Live ones aren't safe around him. (No mistakes)	1	2	3	4	5	☐
35. (11-13) W	1) 2) ③ 4)	I wish you hadn't given those kids such loud whistles. They has blown them all day. (No mistakes)	1	2	3	4	5	☐
36. (12-14) B	1) ② 3) 4)	You should have asked for help. You hadn't ought to lift anything that heavy yourself. (No mistakes)	1	2	3	4	5	☐
37. (13-14) W	1) ② 3) 4)	Mom and Dad are already at the park. Time we get there they will have left. (No mistakes)	1	2	3	4	5	☐
38. (14) B	① 2) 3) 4)	Don't just set on your hands. Help me move this sofa against that wall. (No mistakes)	1	2	3	4	5	☐

REFERENCES

Merz, W. R. Test fairness and test bias: A review of procedures. In, M. J. Wargo & D. R. Green (Eds.), Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. Monterey: CTB/McGraw-Hill, 1978.

Green, D. R. Problems of achievement tests in program evaluation. In M. J. Wargo & D. R. Green (Eds.), Achievement Testing of Disadvantaged and Minority students for Educational Program Evaluation. Monterey: CTB/McGraw-Hill, 1978.